International Tables for Crystallography (2006). Vol. G, Chapter 1.1, pp. 2–10.

1.1. Genesis of the Crystallographic Information File

By S. R. Hall and B. McMahon

1.1.1. Prologue

Progress in science depends crucially on the ability to find and share theories, observations and the results of experiments. The efficient exchange of information within and across scientific disciplines is therefore of fundamental importance. The rapid growth in the use of computers and networks over the past half century and in the use of the World Wide Web over the past decade have brought remarkable improvements in communications among scientists. Such communications are most effective when there is a common language. The English language has become the standard means of expression of ideas and theories; increasingly, the exchange of data requires the computer equivalent of such a *lingua franca*. It was with considerable foresight that the International Union of Crystallography (IUCr) in 1990 adopted a data-handling approach based on universal file concepts. At that time this was considered to be a radical idea. The approach adopted by the IUCr is known as the Crystallographic Information File (CIF). This volume of *International Tables for Crystallography* describes the CIF approach, the associated definition of CIF data items within dictionaries, and handling procedures, applications and software.

In this opening chapter, we give a historical perspective on the reasons why the CIF approach was adopted and how, over the past decade, CIF applications have evolved.

CIF is the most fully developed and mature of the various universal file approaches available today. It combines flexibility and simplicity of expression with a lean syntax. It has an unsurpassed ability to express ‘hard’ scientific data unambiguously using extensive dictionaries (ontologies) of relevant terms. It has proved to be remarkably well suited to the publication and archiving of small-unit-cell crystallographic structures. What was a radical idea in 1990 has today become the dominant mode of expression of scientific data in this domain.

The CIF data model provided the key to the internal restructuring of data managed by the Protein Data Bank in its transition from an archive to a database. The CIF approach is being tested in an increasing number of domains. In some cases, it may well become as successful as it has been for small-molecule crystallography. In other cases, the syntax will be unsuitable, but yet the conceptual discipline of agreed ontologies will still be required. Here, the experience of developing the CIF dictionaries may be carried across into different file formats and modes of expression.

Nowadays, informatics is a rapidly evolving field, in which everything is obsolete almost as soon as it is created. Yet there is a responsibility on today’s scientists to preserve data and pass them on to the next generation. CIF was developed not only as a data-exchange mechanism, but also as an archival format, and considerable care has been taken over the past decade and more to keep it a stable and smoothly evolving approach. Some points of detail have been modified or superseded in practice. Other changes will necessarily occur as the approach evolves to meet the changing demands of an evolving science. Readers should therefore be aware of the need to consult the IUCr website (http://www.iucr.org/iucr-top/cif) for the latest versions of, or successors to, the data dictionaries and the software packages described in this volume. However, the basic concepts have already been shown to be remarkably effective and durable. This volume should therefore provide an invaluable reference for those working with CIF and related universal file approaches.

The success of any data-exchange approach depends on its efficiency and flexibility. It must cope with the increasing volume and complexity of data generated by the computing ‘information explosion’. This growth challenges conventional criteria for measuring exchange and storage efficiency based on high data-compression factors. Today’s fast, cheap magnetic and chip technologies make bulk volume a secondary consideration compared with extensibility and portability of data-management processes. Most importantly, improvements in computing technology continue to generate new approaches to harnessing semantic information contained within data collections and to promoting new strategies for knowledge management.

The basis for an efficient information-exchange process is mutually agreed rules for the supplier and the receiver, *i.e.* the establishment of an exchange protocol. This protocol needs to be established at several levels. At the first level, there must be predetermined ways that data (*i.e.* numbers, characters or text) are arranged in the storage medium. These are the organizational rules that define the syntax or the format of data. There must also be a clear understanding of the meaning of individual data items so that they can be correctly identified, accessed and reused by others. At an even higher level, a protocol may also provide rules for expressing the relationships between the data, as this can lead to automatic processes for validating and applying the data values.

These higher levels provide the *semantic* knowledge needed for the rigorous identification and validation of transmitted and stored data. One may consider the analogy of reading this paragraph in English. To do this we must first be able to recognize the individual words, comprehend their meaning (if necessary with the use of an English dictionary) and understand all of this within the context of sentence construction. As with data, the arrangement of component words is based on a predetermined grammatical syntax, and their individual meaning (as defined in a dictionary or elsewhere), coupled with their contextual function (as nouns, verbs, adjectives *etc.*), leads to the full comprehension of a word sequence as semantic information.

1.1.2. Past approaches to data exchange

The crystallographic community, along with many other scientific disciplines, has long adhered to the philosophy that experimental data and results should be routinely archived to facilitate long-term knowledge retention and access. An early approach to this, recommended by IUCr and other journals, was for authors to deposit data as hard copy (*i.e.* ink on paper) with the British Library Lending Division. Retaining good records is fundamental to reproducing

Affiliations: Sydney R. Hall, School of Biomedical and Chemical Sciences, University of Western Australia, Crawley, Perth, WA 6009, Australia; Brian McMahon, International Union of Crystallography, 5 Abbey Square, Chester CH1 2HU, England.

scientific results. However, the sheer volume of diffraction data needed to repeat a crystallographic study precludes these from publication, and has led in the past to relatively *ad hoc* procedures for depositing supplementary data in local or centralized archives. Typically in the past, only the crystal and structure model parameters were published in the refereed paper and the underpinning diffraction information had to be archived elsewhere. Because the archived data were usually stored as paper in various unregulated formats, considerable information about the experiment and structure-refinement parameters was never retained. Moreover, the archiving of supplementary data *via* postal services was very slow and labour-intensive; equally, the recovery of deposited data was difficult, with information supplied as either a photocopy of the original deposition or an image taken from a microfiche.

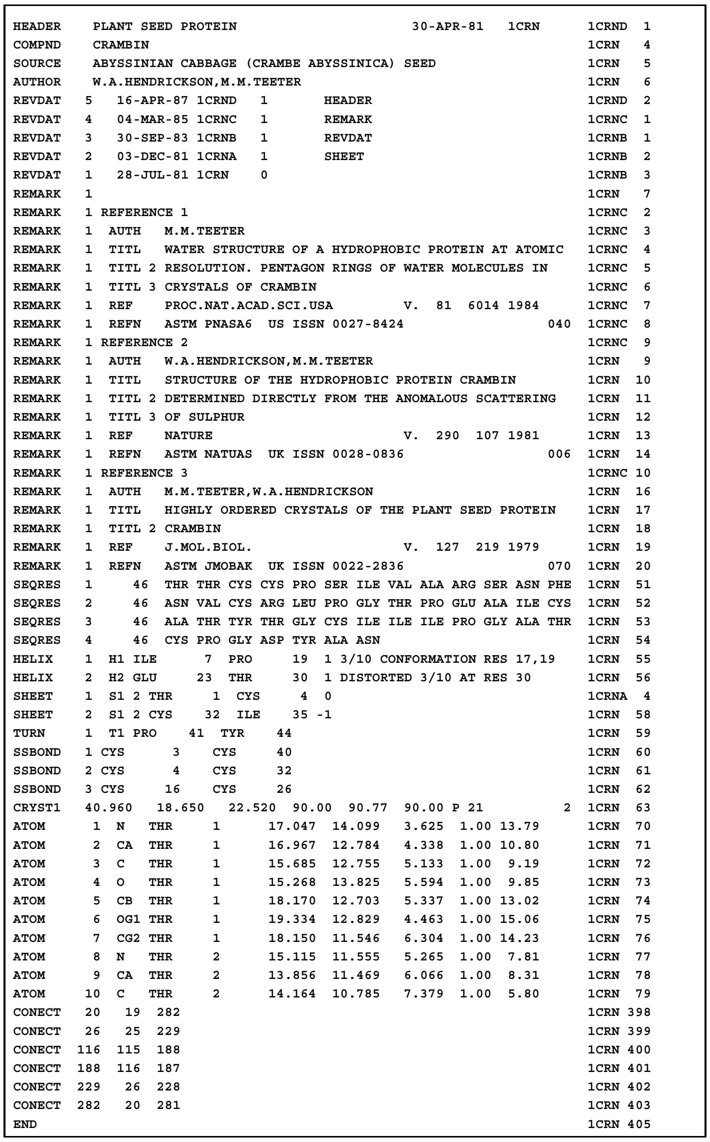
Prior to 1970, when less than 9000 structures were deposited with the Cambridge Crystallographic Data Centre, data sets were still small enough to make these deposition and retrieval approaches feasible, albeit tedious. Even so, records show that very few archived data sets were ever retrieved for later use. The rationale of data storage changed radically in the 1980s. The increasing role of computers, automatic diffractometers and phase-solving direct methods in crystallographic studies led to a rapid acceleration in the number and size of structures determined and published. This was the period when fast minicomputers became affordable for laboratories, and the consequent demand for the electronic storage and exchange of information grew exponentially. Typical data-archival practices changed from using paper to magnetic tapes, as these now became the least expensive and most efficient means of storing data.

1.1.3. Card-image formats

Although the interchange of scientific information depends implicitly on an agreed data format, it remains independent of whether the transmission medium is paper tape, punched card, magnetic tape, computer chip or the Internet. Crystallography has employed countless data-exchange approaches and formats over the past 60 years. Prior to the advent of computers, the standard approach involved the exchange of typed tables of coordinates and structure factors with descriptive headers. In the 1950s and 1960s, as computers became the dominant generators of data, the transfer of data between laboratories was still relatively uncommon. When it was necessary, the Hollerith card formats of commonly used programs, such as *ORFLS* (Busing *et al.*, 1962) and *XRAY* (Stewart, 1963), usually sufficed. Even when magnetic tape drives became common and were standardized (mainly to the 1/2-inch 2400-foot reel), the 80-column 'card-image' formats of these programs remained the most popular data exchange and deposition approach.

As the storage and transporting of electronic data became easier and cheaper, structural information was increasingly deposited directly in databases such as the Cambridge Structural Database (CSD; Allen, 2002) and the Protein Data Bank (PDB; Bernstein *et al.*, 1977). The CSD and PDB simplified these depositions by using standard layouts such as the ASER, BCCAB and PDB formats. Both the PDB and CSD used, and indeed still use as a backup deposition mode, fixed formats with 80-character records and identifier codes. Examples of these format styles are shown in Figs. 1.1.3.1 and 1.1.3.2.

The card-image approach, involving a rigid preordained syntax, survived for more than two decades because it was simple, and the suite of data types used to describe crystal structures remained relatively static.



```
HEADER    PLANT SEED PROTEIN                      30-APR-81   1CRN      1CRND  1
COMPND    CRAMBIN                                                       1CRN   4
SOURCE    ABYSSINIAN CABBAGE (CRAMBE ABYSSINICA) SEED                   1CRN   5
AUTHOR    W.A.HENDRICKSON,M.M.TEETER                                    1CRN   6
REVDAT   5   16-APR-87 1CRND   1       HEADER                           1CRND  2
REVDAT   4   04-MAR-85 1CRNC   1       REMARK                           1CRNC  1
REVDAT   3   30-SEP-83 1CRNB   1       REVDAT                           1CRNB  1
REVDAT   2   03-DEC-81 1CRNA   1       SHEET                            1CRNB  2
REVDAT   1   28-JUL-81 1CRN    0                                        1CRNB  3
REMARK   1                                                              1CRN   7
REMARK   1 REFERENCE 1                                                  1CRNC  2
REMARK   1  AUTH   M.M.TEETER                                           1CRNC  3
REMARK   1  TITL   WATER STRUCTURE OF A HYDROPHOBIC PROTEIN AT ATOMIC   1CRNC  4
REMARK   1  TITL 2 RESOLUTION. PENTAGON RINGS OF WATER MOLECULES IN     1CRNC  5
REMARK   1  TITL 3 CRYSTALS OF CRAMBIN                                  1CRNC  6
REMARK   1  REF    PROC.NAT.ACAD.SCI.USA        V.  81  6014 1984      1CRNC  7
REMARK   1  REFN   ASTM PNASA6  US ISSN 0027-8424                 040  1CRNC  8
REMARK   1 REFERENCE 2                                                  1CRNC  9
REMARK   1  AUTH   W.A.HENDRICKSON,M.M.TEETER                           1CRN   9
REMARK   1  TITL   STRUCTURE OF THE HYDROPHOBIC PROTEIN CRAMBIN         1CRN  10
REMARK   1  TITL 2 DETERMINED DIRECTLY FROM THE ANOMALOUS SCATTERING    1CRN  11
REMARK   1  TITL 3 OF SULPHUR                                           1CRN  12
REMARK   1  REF    NATURE                       V. 290  107 1981       1CRN  13
REMARK   1  REFN   ASTM NATUAS  UK ISSN 0028-0836                 006  1CRN  14
REMARK   1 REFERENCE 3                                                  1CRNC 10
REMARK   1  AUTH   M.M.TEETER,W.A.HENDRICKSON                           1CRN  16
REMARK   1  TITL   HIGHLY ORDERED CRYSTALS OF THE PLANT SEED PROTEIN    1CRN  17
REMARK   1  TITL 2 CRAMBIN                                              1CRN  18
REMARK   1  REF    J.MOL.BIOL.                  V. 127  219 1979       1CRN  19
REMARK   1  REFN   ASTM JMOBAK  UK ISSN 0022-2836                 070  1CRN  20
SEQRES   1     46  THR THR CYS CYS PRO SER ILE VAL ALA ARG SER ASN PHE  1CRN  51
SEQRES   2     46  ASN VAL CYS ARG LEU PRO GLY THR PRO GLU ALA ILE CYS  1CRN  52
SEQRES   3     46  ALA THR TYR THR GLY CYS ILE ILE ILE PRO GLY ALA THR  1CRN  53
SEQRES   4     46  CYS PRO GLY ASP TYR ALA ASN                          1CRN  54
HELIX    1  H1 ILE      7  PRO     19  1 3/10 CONFORMATION RES 17,19    1CRN  55
HELIX    2  H2 GLU     23  THR     30  1 DISTORTED 3/10 AT RES 30       1CRN  56
SHEET    1  S1 2 THR     1  CYS     4  0                                1CRNA  4
SHEET    2  S1 2 CYS    32  ILE    35 -1                                1CRN  58
TURN     1  T1 PRO    41  TYR    44                                     1CRN  59
SSBOND   1 CYS      3    CYS     40                                     1CRN  60
SSBOND   2 CYS      4    CYS     32                                     1CRN  61
SSBOND   3 CYS     16    CYS     26                                     1CRN  62
CRYST1   40.960   18.650   22.520  90.00  90.77  90.00 P 21          2  1CRN  63
ATOM      1  N   THR     1      17.047  14.099   3.625  1.00 13.79      1CRN  70
ATOM      2  CA  THR     1      16.967  12.784   4.338  1.00 10.80      1CRN  71
ATOM      3  C   THR     1      15.685  12.755   5.133  1.00  9.19      1CRN  72
ATOM      4  O   THR     1      15.268  13.825   5.594  1.00  9.85      1CRN  73
ATOM      5  CB  THR     1      18.170  12.703   5.337  1.00 13.02      1CRN  74
ATOM      6  OG1 THR     1      19.334  12.829   4.463  1.00 15.06      1CRN  75
ATOM      7  CG2 THR     1      18.150  11.546   6.304  1.00 14.23      1CRN  76
ATOM      8  N   THR     2      15.115  11.555   5.265  1.00  7.81      1CRN  77
ATOM      9  CA  THR     2      13.856  11.469   6.066  1.00  8.31      1CRN  78
ATOM     10  C   THR     2      14.164  10.785   7.379  1.00  5.80      1CRN  79
CONECT   20   19  282                                                   1CRN 398
CONECT   26   25  229                                                   1CRN 399
CONECT  116  115  188                                                   1CRN 400
CONECT  188  116  187                                                   1CRN 401
CONECT  229   26  228                                                   1CRN 402
CONECT  282   20  281                                                   1CRN 403
END                                                                     1CRN 405
```

Fig. 1.1.3.1. An abbreviated example of a PDB format file.

1.1.4. The Standard Crystallographic File Structure (SCFS)

By the 1980s, the many different fixed formats used to exchange data electronically had become a significant complication for journals and databases. Because of this, the IUCr Commissions for Crystallographic Data and Computing formed a joint Working Party which was asked to recommend a standard format for the exchange and retention of crystallographic data. They proposed a partially fixed format in which key words on each line identified blocks of data containing items in a specific order. This format was the Standard Crystallographic File Structure (Brown, 1988). An example of an SCFS file is shown in Fig. 1.1.4.1.

The effectiveness of the SCFS format approach was curtailed because its release coincided with the arrival of powerful minicomputers, such as the VAX780, in crystallographic laboratories. This led to a period of enormous change in crystallographic computing, in which new data types and file formats proliferated. It was also a time when automatic diffractometers became standard equipment in laboratories and the development of new crystallographic software packages flourished. The fixed-format design of the SCFS was unable to adapt easily to these continually changing data requirements, and this eventually led to a proliferation of SCFS versions.

```
?BAWGEL
#ADATE 820705
#COMPND bis(Benzene)-chromium bromide
#FORMUL C12 H12 Cr1 1+,Br1 1-
#AUTHOR A.L.Spek,A.J.M.Duisenberg
#JRNL 189,10,1531,1981
#CREF msdb 14.74.003 nbsid 532193 batch 53 cdvol 6
#CLASS 1/74
#SYSCAT sys O cat 3
#CONN El= Cr 2 Br 14 V= 1 2 Ch= + 2 Ch= - 14
Res= Plot= 1 B= 5 1-3 1-4 3-6 4-7 5-8 5-9 6-10 7-10 8-11 9-12 11-13 12-13
 B= 9 1-2-5
Res= Plot= 1 14
#DIAGRAM
   469   151   374   202   469   248   554   103   272   248   556   299
   640   152   186   296   272   151   640   250   101   247   185   100
   100   149   100    50     0     0     0     0     0     0     0     0
#CELL a 9.753(6) b 9.316(3) c 11.941(8) z 4 cent 1 sg Fmmm
#SYMM  x,y,z
x,1/2+y,1/2+z
1/2+x,y,1/2+z
1/2+x,1/2+y,z
-x,y,z
-x,1/2+y,1/2+z
1/2-x,y,1/2+z
1/2-x,1/2+y,z
x,-y,z
x,1/2-y,1/2+z
1/2+x,-y,1/2+z
1/2+x,1/2-y,z
-x,-y,z
-x,1/2-y,1/2+z
1/2-x,-y,1/2+z
1/2-x,1/2-y,z
#DENSITY dx 1.764
#UNIS int 3 sigcc 3
#RFACT R= 0.0540.
#RADIUS C  0.68 H  0.23 Br 1.21 Cr 1.35
#TOLER 0.40
#ATOM Cr1 0.0 0.0 0.0
Br1 0.0 0.0 0.50000
C1 0.06900 0.12800 0.13400
C2 0.13900 0.0 0.13400
H1 0.09300 0.20400 0.12500
H2 0.19800 0.0 0.13000
#BOND Cr1 C1 2.100
Cr1 C2 2.090
C1 C1* 1.340
C1 C2 1.370
C1 H1 0.760
C2 H2 0.580
#MDATE  901205
#END
```

Fig. 1.1.3.2. An example of a CSD BCCAB format file.

```
TITLE                                                                       00001
*p6122                   CIFIO  05-Mar87        p6122                       00002
                                                                            00003
SG NAME                                                                     00004
 LATT     NP                                                                00005
 SYST     HEXAGONAL                                                         00006
 BRAV     HEXAGONAL                                                         00007
 HALL     p_61_2_(0_0_-1)                                                   00008
 HERM     p_61_2_2                                                          00009
*EOS                                                                        00010
                                                                            00011
SYMMETRY R11 2 3     T1      R21 2 3     T2      R31 2 3     T3             00012
 SYOP      1 0 0  .0000000     0 1 0  .0000000     0 0 1  .0000000   1      00013
 SYOP     -1 0 0  .0000000     0-1 0  .0000000     0 0 1  .5000000   2      00014
 SYOP      0-1 0  .0000000    -1 0 0  .0000000     0 0-1  .8333330   3      00015
 SYOP      0 1 0  .0000000     1 0 0  .0000000     0 0-1  .3333330   4      00016
 SYOP      1-1 0  .0000000     0-1 0  .0000000     0 0-1  .0000000   5      00017
 SYOP     -1 1 0  .0000000     0 1 0  .0000000     0 0-1  .5000000   6      00018
 SYOP      1 0 0  .0000000     1-1 0  .0000000     0 0-1  .1666670   7      00019
 SYOP     -1 0 0  .0000000    -1 1 0  .0000000     0 0-1  .6666670   8      00020
 SYOP      0-1 0  .0000000     1-1 0  .0000000     0 0 1  .3333330   9      00021
 SYOP      0 1 0  .0000000    -1 1 0  .0000000     0 0 1  .8333330  10      00022
 SYOP      1-1 0  .0000000     1 0 0  .0000000     0 0 1  .1666670  11      00023
 SYOP     -1 1 0  .0000000    -1 0 0  .0000000     0 0 1  .6666670  12      00024
*EOS                                                                        00025
                                                                            00026
FORMULA  EL  NUM                                                            00036
 FORL    s     .5000o    .5000c    1.0000                                   00037
*EOS                                                                        00038
                                                                            00039
CONDITIONS                                                                  00040
 CELLPAREX      .7107     566.00                                            00041
 INT PAREX      .7107     566.00      .147       .681        92             00042
 HKL PARE           0          2         0          4         0         12  00043
 EQIVPARE          92        525                                            00044
*EOS                                                                        00045
                                                                            00046
ATOMS    NAME    X U11   Y U22    Z U33   U U12   P U13     U23  MUL  AT  DT 00052
 UALL                .03500                                                 00053
 ATCO    s           .20140  .79860  .91667         1.00000          6  s   200054
 ATCE    s           .00040  .00040  .00000          .00000                 00055
 UIJ     s           .04100  .04100  .01000  .02500 -.00400 -.00400          00056
 UIJE    s           .00800  .00800  .00700  .00700  .00500  .00500          00057
 ATCO    o           .50100  .50100  .66667         1.00000          6  o   200058
 ATCE    o           .00300  .00300  .00000          .00000                 00059
 UIJ     o           .08900  .08900  .09000  .06300  .00900 -.00900          00060
 UIJE    o           .01800  .01800  .02000  .01900  .00800  .00800          00061
 ATCO    c 1         .49200  .09700  .03780         1.00000         12  c   200062
 ATCE    c 1         .00300  .00300  .00110          .00000                 00063
 UIJ     c 1         .03170  .03170  .03170  .01585  .00000  .00000          00064
 UIJE    c 1         .00000  .00000  .00000  .00000  .00000  .00000          00065
*EOS                                                                        00066
                                                                            00067
CELL DIMENSIONS  A          B         C       ALPHA     BETA      GAMMA   Z 00068
 CELLPARE      8.5300    8.5300   20.3700   90.0000   90.0000  120.0000 12.000069
 ERRSPARE       .0100     .0100     .0100     .0100     .0100     .0100     00070
 VOL PARE    1283.571    3.0775               .5595                         00071
 PHYSPARE                         566.0000                                  00072
*EOS                                                                        00073
                                                                            00074
END                                                                         00081
```

Fig. 1.1.4.1. An abbreviated example of a Standard Crystallographic File Structure (SCFS) format file.

1.1.5. The impact of networking on crystallography

The growth in power of individual minicomputers inevitably helped the development of computational techniques in crystallography. Yet perhaps a more profound development was networking – the ability to exchange electronic data directly between computers. The laborious procedures for transferring information by manual keystroke or exchange of card decks and magnetic tapes were replaced by error-free programmatic procedures. Initially, data could flow easily between computers in the same laboratory; then colleagues could exchange data between scientific departments on the same campus; and before long experimental results, programs and general communications were flowing freely across national and international networks.

During the 1960s, networking was *ad hoc* and proprietary, and rarely extended effectively outside the laboratory. By the 1970s, however, a few standard networking protocols were becoming established. These included uucp, which promoted the growth of dial-up networking between university campuses, and TCP/IP, the transport protocol underlying the ARPANET, that would eventually give rise to the dominant Internet with which we are familiar today. The potential for improving the practice of crystallography through the ease of communications afforded by computer networks was very clear. However, the technology was still costly and required much effort and expertise to implement. Even towards the end of the decade, a meeting of protein crystallographers concluded (Freer & Stewart, 1979) that

> The possibility and usefulness of establishing a computer network for communication among crystallographic laboratories was discussed. The implications for rapid updating and the ease with which programs and data could be transfered among the groups was clearly recognized by all present; however, immediate implementation of a network was not deemed practical by a majority of the participants.

By the mid-1980s, the establishment of a global computer network was well under way. There was still some diversity of transmission protocols on an international scale: uucp, BITNET and X.25 Coloured Book protocols were still competing with TCP/IP, so that communication between different networks had to be managed through gateways. Nevertheless, there was sufficient standardization that it was feasible to communicate with colleagues world-wide by e-mail, to transfer files by ftp and to log in to remote computers by telnet. E-mail, in particular, allowed for the rapid transmission of ASCII text in an arbitrary format. In many respects, this established a goal for other exchange formats to achieve. The establishment of anonymous ftp sites permitted the free exchange of software and data to any user; no special privileges on the host computer were needed. Such availability of electronic information fitted particularly well with the scientific ethic of open exchange of information.

By the early 1990s, TCP/IP and the Internet dominated international networking. The practices of open exchange of information were developed through a number of initiatives. *Gopher* (Anklesaria *et al.*, 1993) provided a general mechanism to access material categorized and published from a computerized information store. *WAIS* (Kahle, 1991), a wide-area information server application designed to service queries conforming to the Z39.50 information retrieval protocol (ANSI/NISO, 1995), provided an effective distributed search engine. The rapid proliferation of new techniques for searching and retrieving information from the Internet was capped in the mid-1990s by the rapid growth in sites implementing hypertext servers (Berners-Lee, 1989). The World Wide Web had become a reality.

The increasing access to global network facilities during the 1980s led to a growing interest among crystallographers in submitting manuscripts to journals electronically, especially for small-molecule structure studies. The Australian delegation at the 1987 General Assembly of the XIVth IUCr Congress in Perth proposed that IUCr journals (specifically *Acta Crystallographica*) should be able to accept manuscripts submitted electronically. It was argued that this would reduce effort on the part of the authors and the journal office in preparation and transcription of manuscripts, and as a consequence reduce costs and transcription errors and simplify data-validation approaches. The acceptance of this General Assembly resolution led to the creation of a Working Party on Crystallographic Information (WPCI), which had as its mandate the investigation of possible approaches to enable the electronic submission of crystallographic research publications.

1.1.6. The Working Party on Crystallographic Information (WPCI)

The WPCI first convened at the 1988 ECM11 conference in Vienna. In the discussions leading up to this meeting, it was widely appreciated that electronic submissions to journals and databases involved data types (*e.g.* manuscript texts, graphical diagrams, the full suite of crystallographic data) that were beyond those accommodated within the SCFS format promoted by the IUCr Data and Computing Commissions. Consequently, it was suggested at the Vienna meeting that a general and extensible universal file approach, similar to the recently developed Self-defining Text Archive and Retrieval (STAR) File format (Hall, 1991; Hall & Spadaccini, 1994), might also be suitable for crystallographic data applications.

At this meeting, it was decided that a WPCI working group, led by Syd Hall, should investigate the development of a universal file protocol that would be suitable for crystallographic data needs. Other universal formats existed, such as ASN.1 (ISO, 2002), which was used for data communications, JCAMP-DX (McDonald & Wilks, 1988), which was used for archiving infrared spectra, and the Standard Molecular Data (SMD) format (Barnard, 1990), which was used for the global exchange of chemical structure data. These were considered relatively inefficient for expressing the repetitive data lists commonly used in crystallography. The working group eventually proposed a Crystallographic Information File (CIF) format which had a syntax similar to, but simpler than, the STAR File. Of particular importance because of the rapid changes taking place with data types, the CIF approach provided a very flexible and extensible file structure in which any type of text or numerical data could be arranged in any order. The typical data structure of a CIF is illustrated in Fig. 1.1.6.1, using the same data as presented in the PDB file of Fig. 1.1.3.1. Similarly, Fig. 1.1.6.2 shows the data in the BCCAB file of Fig. 1.1.3.2 in CIF format.

```
data_crambin
_entry.id                          1CRN

_audit.creation_date               1993-04-21
_audit.creation_method             'manual editing of PDB entry'
_audit.update_record
; 1993-04-21 Original PDB entry history recorded here for completeness.
        30-apr-81 deposition.
        28-jul-81 1crn    0
        03-dec-81 correct residue number on strand 1 of sheet s1.
        30-sep-83 insert revdat records
        04-mar-85 insert new publication as reference and renumber
        16-apr-87 change deposition date from 31-apr-81 to 30-apr-81.
;
loop_
    _struct.entry_id
    _struct.title
    1CRN  'Crambin from Abyssinian cabbage (Crambe abyssinica) seed'

loop_
    _citation.id
    _citation.year
    _citation.journal_abbrev
    _citation.journal_volume
    _citation.page_first
    _citation_journal_id_ASTM
    _citation_journal_id_ISSN
    _citation_title
    primary   1984  Biochemistry  23  6796
            ?      0006-2960
;   Raman spectroscopy of homologous plant toxins: crambin and alpha 1- and
    beta-purothionin secondary structures, disulfide conformation, and
    tyrosine environment
;
    1     1984  Proc.Nat.Acad.Sci.USA  81   6014
            pnasa6   0027-8424
;   Water structure of a hydrophobic protein at atomic resolution. Pentagon
    rings of water molecules in crystals of crambin
;
    2     1981  Nature                 280  107
            natuas   0028-0836
;   Structure of the hydrophobic protein crambin determined directly
    from the anomalous scattering of sulphur
:
loop_
    _citation_author.citation_id
    _citation_author.name
    primary   'Williams, R.W.'        primary   'Teeter, M.M.'
    1         'Teeter, M.M.'
    2         'Hendrickson, W.A.'     2         'Teeter, M.M.'

loop_
    _entity.id
    _entity.type
    _entity.details
    1     polymer       'Protein chain: *'
    2     non-polymer   'het group EOH'

loop_
    _entity_poly_seq.entity_id
    _entity_poly_seq.num
    _entity_poly_seq.mon_id
       1   1 THR    1   2 THR    1   3 CYS    1   4 CYS    1   5 PRO
       1   6 SER    1   7 ILE    1   8 VAL    1   9 ALA    1  10 ARG
       1  11 SER    1  12 ASN    1  13 PHE    1  14 ASN    1  15 VAL
#.........................................sequence data omitted for brevity

_cell.length_a                     40.960
_cell.length_b                     18.650
_cell.length_c                     22.520
_cell.angle_alpha                  90.00
_cell.angle_beta                   90.77
_cell.angle_gamma                  90.00
_symmetry.space_group_name_H-M 'P 1 21 1'

loop_
    _atom_type.symbol
    _atom_type.description
    _atom_type.number_in_cell
    C carbon 404  N nitrogen 112  O oxygen 128  S sulfur 12  H hydrogen ?

loop_
    _atom_site.label_seq_id
    _atom_site.type_symbol
    _atom_site.label_atom_id
    _atom_site.label_comp_id
    _atom_site.label_asym_id
    _atom_site.auth_seq_id
    _atom_site.label_alt_id
    _atom_site.Cartn_x
    _atom_site.Cartn_y
    _atom_site.Cartn_z
    _atom_site.occupancy
    _atom_site.B_iso_or_equiv
    _atom_site.label_entity_id
    _atom_site.id
      1  N  N    THR *  1  .  17.047  14.099  3.625  1.00 13.79  1   1
      1  C  CA   THR *  1  .  16.967  12.784  4.338  1.00 10.80  1   2
      1  C  C    THR *  1  .  15.685  12.755  5.133  1.00  9.19  1   3
      1  O  O    THR *  1  .  15.268  13.825  5.594  1.00  9.85  1   4
      1  C  CB   THR *  1  .  18.170  12.703  5.337  1.00 13.02  1   5
      1  O  OG1  THR *  1  .  19.334  12.829  4.463  1.00 15.06  1   6
      1  C  CG2  THR *  1  .  18.150  11.546  6.304  1.00 14.23  1   7
#.........................................atom-site data omitted for brevity
```

Fig. 1.1.6.1. Example 1 of a CIF (using the same data as shown in Fig. 1.1.3.1).

```
data_BAWGEL
_audit_creation_date              93-05-24
_audit_creation_method            manual_conversion_of_ccdc_file
_audit_update_record
;  82-07-05  CCDC entry created from journal data
                      A.L.Spek,A.J.M.Duisenberg (1981) 189,10,1531
   93-05-21  Received file from Owen Johnson, CCDC.
   93-05-24  Initial conversion of the file to CIF/MIF format.
;
_chemical_name_systematic         'bis(Benzene)-chromium bromide'
_chemical_formula_moiety          'C12 H12 Cr1 1+,Br1 1-'

_cell_length_a                    9.735(6)
_cell_length_b                    9.316(3)
_cell_length_c                    11.941(8)
_cell_angle_alpha                 90
_cell_angle_beta                  90
_cell_angle_gamma                 90
_cell_formula_units_Z             4

_symmetry_space_group_name_H-M    Fmmm

loop_
    _symmetry_equiv_pos_as_xyz
        x,y,z x,1/2+y,1/2+z 1/2+x,y,1/2+z 1/2+x,1/2+y,z -x,y,z -x,1/2+y,1/2+z
        1/2-x,y,1/2+z 1/2-x,1/2+y,z  x,-y,z x,1/2-y,1/2+z 1/2+x,-y,1/2+z
        1/2+x,1/2-y,z -x,-y,z -x,1/2-y,1/2+z 1/2-x,-y,1/2+z 1/2-x,1/2-y,z

loop_
    _atom_type_symbol
    _atom_type_radius_bond
                          C 0.68   H 0.23   Br 1.21   Cr 1.35

_exptl_crystal_density_meas       1.764
_refine_ls_R_factor_obs           0.0540
_reflns_observed_criterion        3sigma(I)

loop_
    _atom_site_label
    _atom_site_fract_x
    _atom_site_fract_y
    _atom_site_fract_z
                     Cr1 0.0     0.0     0.0
                     Br1 0.0     0.0     0.50000
                     C1  0.06900 0.12800 0.13400
                     C2  0.13900 0.0     0.13400
                     H1  0.09300 0.20400 0.12500
                     H2  0.19800 0.0     0.13000
loop_
    _geom_bond_atom_site_label_1
    _geom_bond_atom_site_label_2
    _geom_bond_distance
    _geom_bond_site_symmetry_1
    _geom_bond_site_symmetry_2

                        Cr1 C1 2.100  .   .
                        Cr1 C2 2.090  .   .
                        C1  C1 1.340  .   5_555
                        C1  C2 1.370  .   .
                        C1  H1 0.760  .   .
                        C2  H2 0.580  .   .
```

Fig. 1.1.6.2. Example 2 of a CIF (using the same data as shown in Fig.1.1.3.2).

1.1.7. The Crystallographic Information File

As outlined in Section 1.1.6, the working group commissioned by the WPCI set out to establish an exchange protocol suitable for submitting crystallographic data to journals and databases, and this resulted in the development of the CIF syntax. At the same time, the group was also asked to form a list of those data items considered to be essential in a manuscript submitted to *Acta Crystallographica*. The data items originally recommended are listed in Fig. 1.1.7.1.

The syntax of a CIF (a detailed description is given in Chapter 2.2) was intentionally a simple subset of the STAR File syntax (see Chapter 2.1 for details). This simplification was considered important for its easy implementation in existing crystallographic software packages – clearly a primary goal for any format that was to be widely available for submitting data to journals and databases.

A compilation of data names referring to specific quantities or concepts in a crystal-structure determination was drawn up. This compilation included the items already identified as necessary for publication and many more besides. As a list of standard tags intended for unambiguous use, the collection was known from the outset as a *dictionary* of data names.

```
# The following CIF data names encompass the IUCr Journals Commission
# requirements for the reporting of a small-molecule structure in Acta
# Crystallographica, Section C

_chemical_compound_source
_chemical_formula_sum
_chemical_formula_moiety
_chemical_formula_weight
_symmetry_cell_setting
_symmetry_space_group_name_H-M
_cell_length_a
_cell_length_b
_cell_length_c
_cell_angle_alpha
_cell_angle_beta
_cell_angle_gamma
_cell_volume
_cell_formula_units_Z
_exptl_crystal_density_diffrn
_exptl_crystal_density_meas
_exptl_crystal_density_method
_diffrn_radiation_type
_diffrn_radiation_wavelength
_cell_measurement_reflns_used
_cell_measurement_theta_min
_cell_measurement_theta_max
_exptl_absorpt_coefficient_mu
_cell_measurement_temperature
_exptl_crystal_description
_exptl_crystal_colour
_exptl_crystal_size_max
_exptl_crystal_size_mid
_exptl_crystal_size_min
_exptl_crystal_size_rad
_diffrn_measurement_device
_diffrn_measurement_method
_exptl_absorpt_correction_type
_exptl_absorpt_correction_T_min
_exptl_absorpt_correction_T_max
_diffrn_reflns_number
_reflns_number_total
_reflns_number_observed
_reflns_observed_criterion
_diffrn_reflns_av_R_equivalents
_diffrn_reflns_theta_max
_diffrn_reflns_limit_h_min
_diffrn_reflns_limit_h_max
_diffrn_reflns_limit_k_min
_diffrn_reflns_limit_k_max
_diffrn_reflns_limit_l_min
_diffrn_reflns_limit_l_max
_diffrn_standards_number
_diffrn_standards_interval_count
_diffrn_standards_interval_time
_diffrn_standards_decay_%
_refine_ls_structure_factor_coef
_refine_ls_R_factor_obs
_refine_ls_wR_factor_obs
_refine_ls_goodness_of_fit_obs
_refine_ls_number_reflns
_refine_ls_number_parameters
_refine_ls_hydrogen_treatment
_refine_ls_weighting_scheme
_refine_ls_shift/esd_max
_refine_diff_density_max
_refine_diff_density_min
_refine_ls_extinction_method
_refine_ls_extinction_coef
_atom_type_scat_source
_refine_ls_abs_structure_details
_computing_data_collection
_computing_cell_refinement
_computing_data_reduction
_computing_structure_solution
_computing_structure_refinement
_computing_molecular_graphics
_computing_publication_material
_publ_section_experimental
```

Fig. 1.1.7.1. Initial set of data items considered to be essential in a structure report submitted to *Acta Crystallographica Section C*.

The WPCI proposed the CIF format as a standard exchange protocol at the open meetings of the IUCr Commissions on Crystallographic Data and Computing at the 1990 XVth IUCr Congress in Bordeaux. The proposal was accepted and the CIF format was subsequently adopted by the IUCr as the preferred format for data exchange (Hall *et al.*, 1991).

The administration of the CIF standard, including the approval of new data items, is the responsibility of the IUCr Committee for the Maintenance of the CIF Standard (COMCIFS). This committee plays a central role in the coordination of CIF activities, such as the creation of new dictionaries for defining crystallographic data items and the updating of data definitions in existing dictionaries. Chapter 3.1 describes relevant aspects of its role in the commissioning and maintenance cycle. Information about COMCIFS activities and other CIF developments may be obtained from http://www.iucr.org/iucr-top/cif/.

1.1.8. Diversification: the Molecular Information File and dictionary definition language

While the primary thrust of these activities was the development of an exchange mechanism for crystal-structure reports, there was also interest in enriching the description of the chemical properties and behaviour of the compounds under study. Some work was therefore done to broaden the descriptions of bond order that were present in rudimentary form in the core CIF dictionary and to develop more detailed two-dimensional graphical representations of chemical molecules. The result of this work was the Molecular Information File (MIF), described in Chapter 2.4. As with CIF, the specific data items required for MIF were defined in a dictionary.

This work has extended the STAR File approach into chemistry. It is envisaged that later modules could describe spectroscopic data, reaction schemes and much more. Particular requirements of chemical structural databases are the need to query for generic structures, and the need to allow for the labelling and comparison of libraries of substructural components. Both requirements can be met by features of the STAR File, but they are features omitted from CIF. In practice, therefore, data files in MIF format cannot be readily accessed by most crystallographic applications, and the format is at present little used by crystallographers.

An important outcome of the work on MIF was the recognition that attributes of the data items needed for a particular application can be recorded using the same formalism as the data files themselves. This gave rise to the idea of a dictionary definition language (DDL), a set of tags for describing the names and attributes of data items. The dictionaries (the collections of data names for CIF and MIF applications) could then be constructed as STAR Files themselves, with the immediate result that software written to parse data files could equally easily parse the associated dictionaries. Now it became feasible to build into applications the ability to validate data by dynamically reading and interpreting the properties associated with a data tag in an accompanying dictionary.

The idea of a DDL was proposed by Tony Cook during early discussions on MIF and was adopted while the original CIF paper (Hall *et al.*, 1991) was in the press. The original core CIF dictionary was therefore produced with an early version of a DDL that was never fully documented (Fig. 1.1.8.1). Building on early experience with the core dictionary and the technical evolution of MIF, Hall & Cook (1995) worked through several revisions before publishing DDL version 1.4, the stable version described in Chapter 2.5 of this volume. Because of the circumstances in which it was developed, this dictionary definition language is able to accommodate both the flat-file quasi-relational structure of CIF and the more hierarchical multiple-looped data model of MIF.

1.1.9. The macromolecular Crystallographic Information File

The original goal of CIF was the creation of an archive format for the description and results of experiments in small-molecule and inorganic crystallography. The data names and their definitions were embodied in the *core* dictionary, so called because most of the terms in it were considered common to any crystallographic application. In 1990, the IUCr formed a working group, chaired by Paula Fitzgerald, to expand this dictionary to meet the additional requirements of macromolecular crystallography. The resulting expanded dictionary was to be known as the macromolecular CIF (mmCIF) dictionary.

```
##############################################################################
#
#                         DDL Data Name Descriptions
#                         --------------------------
#
# _compliance           The dictionary version in which the item is defined.
#
# _definition           The description of the item.
#
# _enumeration          A permissible value for an item. The value 'unknown'
#                       signals that the item can have any value.
#
# _enumeration_default  The default value for an item if it is not specified
#                       explicitly. 'unknown' means the default is not known.
#
# _enumeration_detail   The description of a permissible value for an item.
#                       Note that the code '.' normally signals a null
#                       or 'not applicable' condition.
#
# _enumeration_range    The range of values for a numerical item. The
#                       construction is 'min:max'. If 'max' is omitted then the
#                       item can have any value greater than or equal to 'min'.
#
# _esd                  Signals whether an estimated standard deviation is
#                       expected to be appended (enclosed within brackets)
#                       to a numerical item. May be 'yes' or 'no'.
#
# _esd_default          The default value for the esd of a numerical item
#                       if a value is not appended.
#
# _example              An example of the item.
#
# _example_detail       A description of the example.
#
# _list                 Signals whether an item is expected to occur in a looped
#                       list. Possible values: 'yes','no' or 'both'.
#
# _list_identifier      Identifies a data item that MUST appear in the list
#                       containing the currently defined data item.
#
# _name                 The data name of the item defined.
#
# _type                 The data type 'numb' or 'char' (latter includes 'text').
#
# _units_extension      The data-name extension code used to specify the units
#                       of a numerical item.
#
# _units_description    A description of the units.
#
# _units_conversion     The method of converting the item into a value based
#                       on the default units. Each conversion number is
#                       preceded by an operator code *, /, +, or - which
#                       indicates how the conversion number is applied.
#
# _update_history       A record of the changes to this file.
#
#-eof-eof-eof-eof-eof-eof-eof-eof-eof-eof-eof-eof-eof-eof-eof-eof-eof-eof-eof
```

Fig. 1.1.8.1. Informal DDL used in the initial version of the core CIF dictionary (now superseded by DDL1.4).

The group's original short-term goal was to fulfil the IUCr mandate of defining data names for an adequate description of a macromolecular crystallographic experiment and its results. Longer-term goals were also determined: to provide sufficient data names so that the experimental section of a structure paper could be written automatically and to facilitate the development of tools so that computer programs could easily interface with CIF data files. A number of informal and formal meetings were held to describe the progress of this project and to solicit community feedback.

An important meeting took place at the University of York in April 1993. The attendees included the mmCIF working group, structural biologists and computer scientists. Vigorous discussion arose on whether the formal structure of the dictionary implemented in the then-current dictionary definition language (DDL1) could deal with the complexity of macromolecular data sets. There were criticisms that the data typing was not strong enough and that there were no formal links expressing relationships between data items. A working group was formed to address these issues, resulting in a second workshop in Tarrytown, New York, in October 1993. The discussions at this second meeting focused on the development of software tools and the requirements of an enhanced DDL. Such a DDL was proposed during a third workshop at the Free University of Brussels in October 1994. This new DDL (DDL2; Chapter 2.6) was designed by John Westbrook and sought to address the various problems identified at the preceding workshops, while retaining compatibility with existing CIF data files.

The macromolecular dictionary was built using DDL2 and was presented as a draft to the community at the American Crystallographic Association meeting in Montreal in July 1995. The draft was subsequently posted on a website and community comment solicited *via* an e-mail discussion list. This provoked lively discussions, leading to continuous correction and updating of the dictionary over an extended period of time. Software for parsing the dictionary and managing mmCIF data sets was developed and was also presented on the website.

In January 1997, the mmCIF dictionary was completed and submitted to COMCIFS for review. In June 1997, version 1.0 was approved by COMCIFS and released (Bourne *et al.*, 1997; Fitzgerald *et al.*, 1996). A workshop was held at Rutgers University in October 1997, at which tutorials were presented to demonstrate the use of the various tools that had been developed.

More than 100 new definitions have been incorporated in version 2 of the mmCIF dictionary, presented in Chapter 4.5.

1.1.10. The Crystallographic Binary File

In 1995, Andy Hammersley approached the IUCr with a proposal to develop an exchange and archival mechanism for image data using the CIF formalism. There was an increasing need to record and exchange two-dimensional images from a growing collection of area detectors and image plates from several manufacturers, each using different and proprietary data-storage formats. The project was encouraged by the IUCr and was developed under a variety of working names. In October 1997, a workshop at Brookhaven National Laboratory, New York, was convened to discuss progress and to coordinate the development of software suitable for this new format.

At the workshop, it became apparent that there was broad consensus to adopt and further develop the working set of data names that had been devised by the working group on this project; it was further decided that the relationships between these data names were best handled by the DDL2 formalism used for mmCIF. While image plates were not the sole preserve of macromolecular crystallography, it was felt that there would be maximum synergy with that community, and that the proposed imgCIF dictionary was most naturally viewed as an extension or companion to the mmCIF dictionary. The adoption of DDL2 would not preclude the generation of DDL1 analogue files if they were found to be necessary in certain applications, but to date such a need has not arisen.

However, on one point the workshop was insistent. For efficient handling of large volumes of image data on the necessary timescales within a large synchrotron research facility, the raw data must be in binary format. It was argued that although ASCII encoding could preserve the information content of a data stream in a fully CIF-compliant format, the consequent overheads in increased file size and data-processing time were unacceptable in environments with a very high throughput of such data sets, even given the high performance of modern computers. Consequently, the Crystallographic Binary File (CBF) was born as an extension to CIF (Chapter 2.3). The CBF may contain binary data, and therefore cannot be considered a CIF (Chapter 2.2). However, except for the representation of the image data, the file retains all the other features of CIF. Information about the experimental apparatus, duration, environmental conditions and operating parameters, together with descriptions of the pixel characteristics of individual frames, are all provided in ASCII character fields tagged by data names that are themselves fully defined in a DDL2-based dictionary (Chapter 3.7).

At first sight, the distinction between CIF and the crystallographic binary file may therefore seem trivial. However, in allowing binary data, the CBF requires greater care in defining the structure and packing by octets of the data, and loses some of the portability of CIF. It also precludes the use of many simple text-based file-manipulation tools. On the other hand, the importing of a stable and well developed tagged information format means that developers do not need to write novel parsers and compatibility with CIFs is easily attained. Indeed, the original idea of a fully compliant image CIF has been retained. By ASCII-encoding the image data in a crystallographic binary file, a fully compliant CIF, known as imgCIF, may be simply generated. One could consider an imgCIF as an archival version and a CBF as a working version of the same information set.

1.1.11. Other extension dictionaries

Since the introduction of the major CIF dictionaries, several other compendia of data names suitable for describing different applications or disciplines within crystallography have been developed. Four of these are described in the current volume.

The powder CIF dictionary (pdCIF; Chapter 3.3) is a supplement to the core dictionary addressing the needs of powder diffractionists. The structural model derived from powder work is familiar to single-crystal small-molecule or inorganic scientists. However, the powder CIF effort had the additional goals of documenting and archiving experimental data. It was always intended that powder CIF be used for communication of completed studies and for data exchange between laboratories. This is frequently done at shared diffraction facilities such as neutron and synchrotron sources. The powder dictionary was written with data from conventional X-ray diffractometers and from synchrotron, continuous-wavelength neutron, time-of-flight neutron and energy-dispersive X-ray instruments in mind.

The modulated-structures dictionary (msCIF; Chapter 3.4) is also considered as a supplement to the core dictionary and is designed to permit the description of incommensurately modulated crystal structures. The project was sponsored by the IUCr Commission on Aperiodic Crystals and was developed in parallel with a standard for the reporting of such structures in the literature (Chapuis *et al.*, 1997).

A small dictionary of terms for reporting accurate electron densities in crystals (rhoCIF; Chapter 3.5) has recently been published as a further supplement to the core dictionary. It has been developed under the sponsorship of the IUCr Commission on Charge, Spin and Momentum Densities.

The symmetry dictionary (symCIF; Chapter 3.8) was developed under the direct sponsorship of COMCIFS with the objective of producing a rigorous set of definitions suitable for the description of crystallographic symmetry. Following its publication, several data names from the symCIF dictionary were incorporated in the latest version of the core dictionary to replace the original informal definitions relating to symmetry. The symCIF dictionary contains most of the data names that would be needed to tabulate space-group-symmetry relationships in the manner of *International Tables for Crystallography* Volume A (2002). It is intended to expand the dictionary to include group–subgroup relations in a later version.

Other dictionaries are also under development, often under the supervision of one of the Commissions of the International Union of Crystallography.

Mention should also be made of the use of STAR Files by the BioMagResBank group at the University of Wisconsin to record NMR structures. This work (Ulrich *et al.*, 1998) endeavours to be complementary to the mmCIF descriptions of structures in the Protein Data Bank.

1.1.12. The broader context: CIF and XML

In the light of more recent data-exchange developments, it will be surprising to newcomers to CIF that more use is not made in crystallography of the extensible markup language XML (W3C, 2001). However, the development of CIF predates XML, and the CIF format can be easily translated to and from suitable XML representations. Most current crystallographic software imports and exports data in CIF format and the use of XML only becomes important in applications that cross the boundary of crystallography and involve interoperability with other scientific domains.

At one time, the antecedent of XML, standard generalized markup language (SGML; ISO, 1986), was considered as a candidate for a crystallographic exchange mechanism. SGML is a highly flexible and extensible system for specifying markup languages, but is extremely general. In the late 1980s, successful SGML implementations stretched the capacity of affordable computers and little accompanying software was available. SGML was at that time a suitable data and document tagging mechanism for large-scale publishers, but was far from appropriate for smaller-scale applications. XML was introduced during the 1990s as a specific SGML markup with a concrete syntax and simplifications that resulted in a lightweight, manageable language, for which robust parsers, editors and other programs could be written and implemented on desktop computers. The consequence has been a very rapid adoption of XML across many disciplines. Parallels may be drawn with the decision to implement CIF as a subset of the more general STAR File.

XML provides the ability to mark up a document or data set with embedded tags. Such tags may indicate a particular typographic representation. More usefully, however, they can reflect the nature or purpose of the information to which they refer. The design of such useful and well structured content tagging (in XML and other formalisms) is referred to in terms of constructing a subject or domain *ontology*. This term is rather poorly defined, but broadly covers the construction for a specific topic area of a controlled vocabulary of terms, the elaboration of relationships between those terms, and rules or constraints governing the use of the terms.

In XML and SGML, document-type definitions (DTDs) and schemas exist as external specifications of the markup tags permitted in a document, their relationships and any optional attributes they might possess. If carefully designed, these have the potential to act as ontologies. A simplification that XML offers over SGML is the ability to construct documents that do not need to conform to a particular schema. This makes it rather easier to develop software for generating and transmitting XML files. However, if diverse applications are to make use of the information content in an XML file, there must be some general way to exchange information about the meaning of the embedded markup, and in practice DTDs or schemas are essential for interoperability between software applications from different sources.

Recent initiatives in chemistry, under the aegis of the International Union of Pure and Applied Chemistry (IUPAC), suggest an active interest in the development of a machine-parsable ontology for chemistry. Building on an existing XML representation of chemical information known as Chemical Markup Language (CML; Murray-Rust & Rzepa, 1999, 2001), IUPAC project groups are mapping out areas of the science for which suitable DTDs and schemas may be constructed that tag relevant chemical content.

In this context, the CIF dictionaries described and annotated at length in this volume provide a sound basis for a machine-parsable ontology for crystallographic data. Given the orderly classification and relationships between tags in a CIF data set, format transformations using XML tags are not at all difficult to achieve. Chapters 5.3 and 5.5 describe tools for converting between mmCIF and XML formats for interchange within the biological structure community.

In the future, we expect to see the growth of XML environments where the full content of the CIF dictionaries is imported for the purpose of tagging crystallographic data embedded in more general documents and data sets. There have already been examples of ontology development using CIF dictionaries; for example, the Object Management Group has developed software classes for middleware in biological computing applications that are modelled on the mmCIF dictionary (Greer, 2000). The use of interactive STAR ontologies is described by Spadaccini *et al.* (2000).

It is possible that as interdisciplinary knowledge-management software systems develop, the exchange of crystallographic data will occur through XML files or other formats. Neverthless, CIF will remain an efficient mechanism for developing detailed data models within crystallography. We can confidently say that, whatever the actual transport format, the intellectual content of the dictionaries in this volume and their subsequent extensions and revisions will continue to underpin the definition and exchange of crystallographic data.

References

Allen, F. H. (2002). *The Cambridge Structural Database: a quarter of a million crystal structures and rising. Acta Cryst.* B**58**, 380–388.

Anklesaria, F., McCahill, M., Lindner, P., Johnson, D., Torrey, D. & Alberti, B. (1993). *The Internet gopher protocol (a distributed document search and retrieval protocol).* RFC 1436. Network Working Group. http://www.ietf.org/rfc/rfc1436.txt.

ANSI/NISO (1995). *Information retrieval (Z39.50): Application service definition and protocol specification.* Z39.50-1995. http://lcweb.loc.gov/z3950/agency/1995doce.html.

Barnard, J. M. (1990). *Draft specification for revised version of the Standard Molecular Data (SMD) format. J. Chem. Inf. Comput. Sci.* **30**, 81–96.

Berners-Lee, T. (1989). *Information management: a proposal.* Internal report. Geneva: CERN. http://www.w3.org/History/1989/proposal-msw.html.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *The Protein Data Bank: a computer-based archival file for macromolecular structures. J. Mol. Biol.* **112**, 535–542.

Bourne, P., Berman, H. M., McMahon, B., Watenpaugh, K. D., Westbrook, J. D. & Fitzgerald, P. M. D. (1997). *Macromolecular Crystallographic Information File. Methods Enzymol.* **277**, 571–590.

Brown, I. D. (1988). *Standard Crystallographic File Structure-87. Acta Cryst.* A**44**, 232.

Busing, W. R., Martin, K. O. & Levy, H. A. (1962). *ORFLS.* Report ORNL-TM-305. Oak Ridge National Laboratory, Tennessee, USA.

Chapuis, G., Farkas-Jahnke, M., Pérez-Mato, J. M., Senechal, M., Steurer, W., Janot, C., Pandey, D. & Yamamoto, A. (1997). *Checklist for the description of incommensurate modulated crystal structures. Report of the International Union of Crystallography Commission on Aperiodic Crystals. Acta Cryst.* A**53**, 95–100.

Fitzgerald, P. M. D., Berman, H., Bourne, P., McMahon, B., Watenpaugh, K. & Westbrook, J. (1996). *The mmCIF dictionary: community review and final approval. Acta Cryst.* A**52** (Suppl.), C575.

Freer, S. T. & Stewart, J. (1979). *Computer programming for protein crystallographic applications, University of California at San Diego, 28-29 November 1978. J. Appl. Cryst.* **12**, 426–427.

Greer, D. S. (2000). *Macromolecular structure RFP response, revised submission.* http://openmms.sdsc.edu/OpenMMS-1.5.1_Std/openmms/docs/specs/lifesci_00-11-01.pdf.

Hall, S. R. (1991). *The STAR file: a new format for electronic data transfer and archiving. J. Chem. Inf. Comput. Sci.* **31**, 326–333.

Hall, S. R., Allen, F. H. & Brown, I. D. (1991). *The Crystallographic Information File (CIF): a new standard archive file for crystallography. Acta Cryst.* A**47**, 655–685.

Hall, S. R. & Cook, A. P. F. (1995). *STAR dictionary definition language: initial specification. J. Chem. Inf. Comput. Sci.* **35**, 819–825.

Hall, S. R. & Spadaccini, N. (1994). *The STAR File: detailed specifications. J. Chem. Inf. Comput. Sci.* **34**, 505–508.

International Tables for Crystallography (2002). Volume A, *Space-group symmetry*, edited by Th. Hahn. Dordrecht: Kluwer Academic Publishers.

ISO (1986). ISO 8879. *Information processing – Text and office systems – Standard Generalized Markup Language (SGML).* Geneva: International Organization for Standardization.

ISO (2002). ISO/IEC 8824-1. *Abstract Syntax Notation One (ASN.1). Specification of basic notation.* Geneva: International Organization for Standardization.

Kahle, B. (1991). *An information system for corporate users: wide area information servers.* Thinking Machines technical report TMC-199. See also *Online*, **15**, 56–60.

McDonald, R. S. & Wilks, P. A. (1988). *JCAMP-DX: a standard form for exchange of infrared spectra in computer readable form. Appl. Spectrosc.* **42**, 151–162.

Murray-Rust, P. & Rzepa, H. S. (1999). *Chemical markup language and XML. Part I. Basic principles. J. Chem. Inf. Comput. Sci.* **39**, 928–942.

Murray-Rust, P. & Rzepa, H. S. (2001). *Chemical markup, XML and the world-wide web. Part II: Information objects and the CMLDOM. J. Chem. Inf. Comput. Sci.* **41**, 1113–1123.

Spadaccini, N., Hall, S. R. & Castleden, I. R. (2000). *Relational expressions in STAR File dictionaries*. *J. Chem. Inf. Comput. Sci.* **40**, 1289–1301.

Stewart, J. (1963). *XRAY63 Crystal Structure Calculations System.* Report TR-64-6 (NSG-398). Computer Science Center, University of Maryland, USA, and Research Computer Center, University of Washington, USA.

Ulrich, E. L. *et al.* (1998). XVIIth Intl Conf. Magn. Res. Biol. Systems. Tokyo, Japan.

W3C (2001). *Extensible Markup Language (XML).* http://www.w3c.org/XML/.